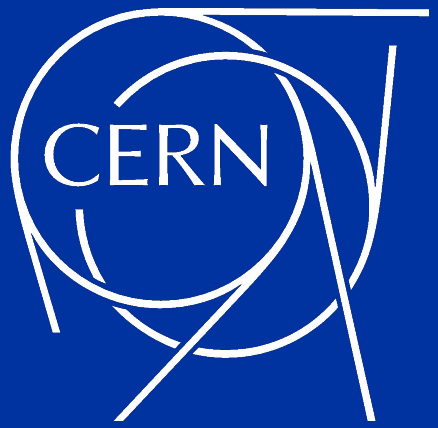


Status of Enabling Archive Metadata for Tapes

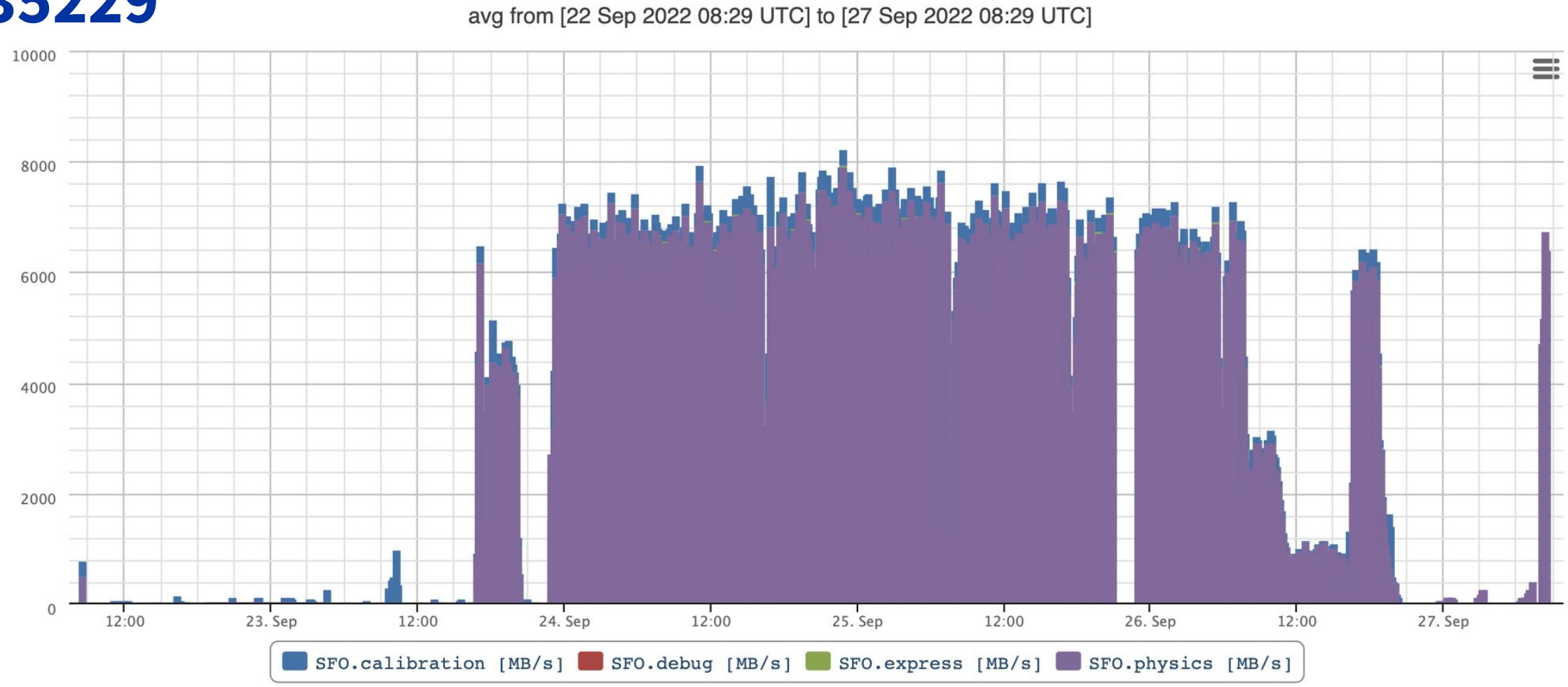
CTA/dCache/FTS/Rucio software implementation feedback

2023-11-10- Julien Leduc from T0 TAPE

Beginning of Run3: *legacy* placement tweaks

- **CTA maps tape family with directory**
  - ***CASTOR legacy***
- **Improving written data placement with the experiments**
  - **Improve per directory tape collocation on tape**
    - **CMS split of MC, 2022 data, 2023 data**
- **Several limitations as**
  - CTA queueing is purely FIFO
    - relies on Tier 0 time collocation for DAQ data
    - delayed retries are mixed with other datasets
  - **CTA directory structures is dictated by experiment namespace**
    - **no directory/file remapping in CTA tape buffer**
  - **Relies exclusively on tape families**
    - **Does not work when tape storage becomes warmer**

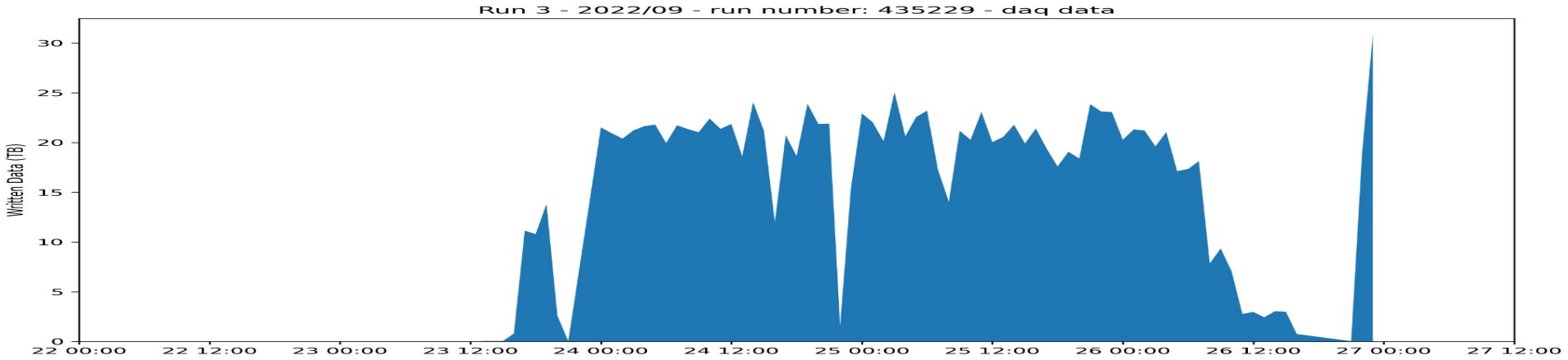
ATLAS Run 435229



ATLAS Run 435229

ATLAS DAQ to T0 latency is 21 minutes

ATLAS Run 435229



	name	wb
10	data22_13p6TeV.435229.physics_MinBias.daq.RAW.	508882192115748
2	data22_13p6TeV.435229.calibration_LArCells.daq...	6029697160416
9	data22_13p6TeV.435229.physics_Main.daq.RAW.	5633626551664
7	data22_13p6TeV.435229.calibration_ZDCCalib.daq...	2060844419964
0	data22_13p6TeV.435229.calibration_AFPCalib.daq...	1095598880584
3	data22_13p6TeV.435229.calibration_LArCellsEmpt...	618731659276
1	data22_13p6TeV.435229.calibration_CostMonitori...	199619777636
6	data22_13p6TeV.435229.calibration_Tile.daq.RAW.	157313259108
5	data22_13p6TeV.435229.calibration_LArPEBDigita...	52435401824
4	data22_13p6TeV.435229.calibration_LArPEB.daq.RAW.	52432509416
8	data22_13p6TeV.435229.calibration_lucid.daq.RAW.	23718811248
11	data22_calib.435229.calibration_MuonAll.daq.RAW.	6442449660

	name	wb
10	data22_13p6TeV.435229.physics_MinBias.daq.RAW.	533995407337536
2	data22_13p6TeV.435229.calibration_LArCells.daq...	6150297188928
9	data22_13p6TeV.435229.physics_Main.daq.RAW.	6010032822852
7	data22_13p6TeV.435229.calibration_ZDCCalib.daq...	2229594391816
0	data22_13p6TeV.435229.calibration_AFPCalib.daq...	1116667277768
3	data22_13p6TeV.435229.calibration_LArCellsEmpt...	576785520352
1	data22_13p6TeV.435229.calibration_CostMonitori...	210124626128
6	data22_13p6TeV.435229.calibration_Tile.daq.RAW.	199264141996
5	data22_13p6TeV.435229.calibration_LArPEBDigita...	83896391044
4	data22_13p6TeV.435229.calibration_LArNoiseBurs...	52431081832
8	data22_13p6TeV.435229.calibration_lucid.daq.RAW.	21562555680
11	data22_calib.435229.calibration_MuonAll.daq.RAW.	6442448020

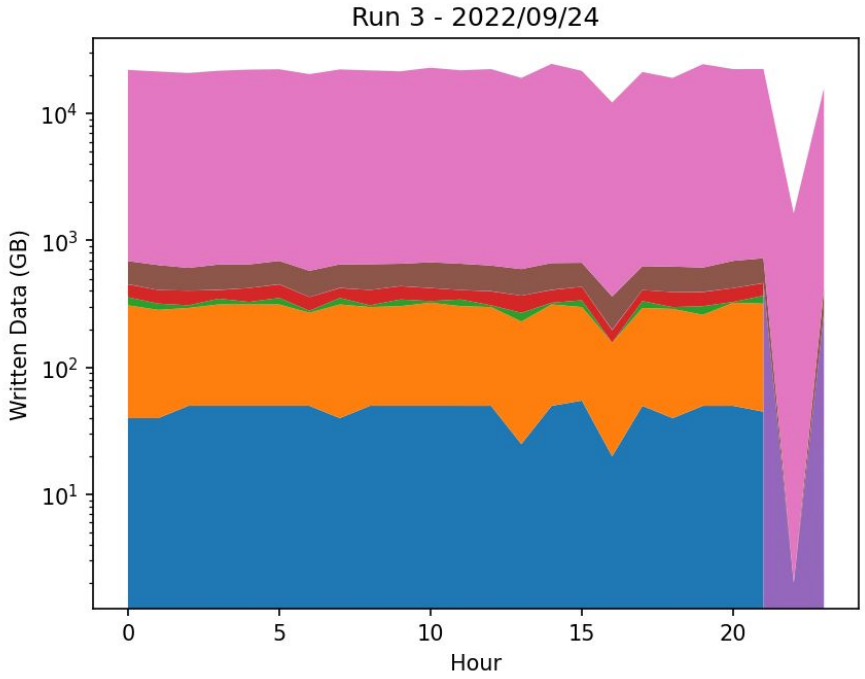
LHC stable beam -> huge datasets

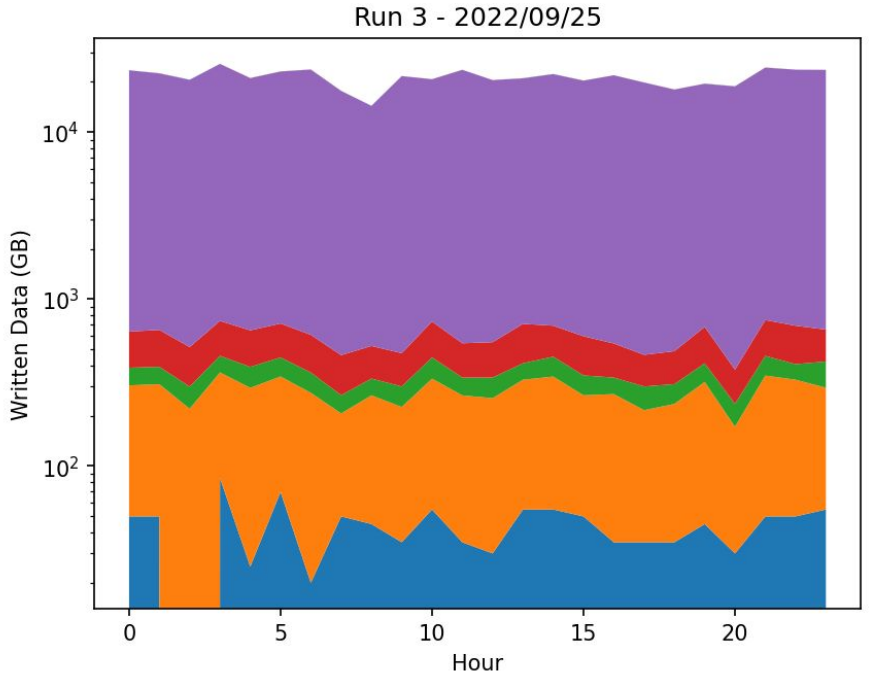
1.8B events, 1.3PB

> 12 other smaller datasets sent in parallel

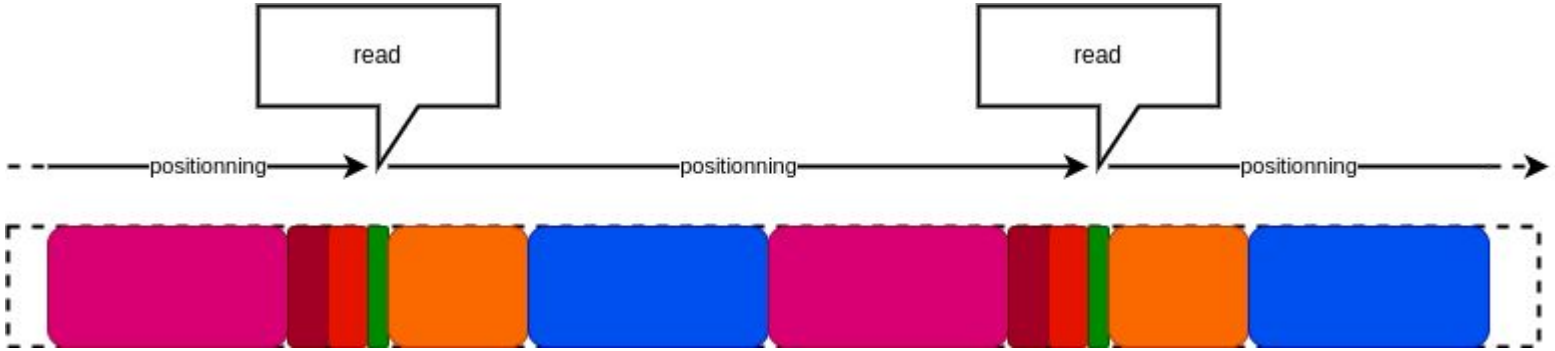
# of parallel datasets sent per run?

BAD FOR CTA DATA PLACEMENT ON TAPE!!





ATLAS Run 435229



- **At DAQ T0 must go to tape ASAP**
  - Cannot be kept for long in buffer
  - Dataset are not complete when they land on tape at T0
- **SAME CONSTRAINTS FOR ALL EXPERIMENTS**
  - LHC and SMEs
- **Collocation hints must be experiment agnostic per file**
  - No experiment specific terminology
  - Not deduced from filename or path

Improve data collocation on tape

Strictly mapping experiment directory structure to tape collocation is not possible:

- **Experiment conventions evolve over time: CMS parking data for example**
  - Initially in a dedicated directory: currently interleaved with DAQ data
- **Multi VO sites cannot deal with experiment specific conventions**
  - At T0 we have already a hard time getting tape throughput and volume expectations…
- **Flat namespace based on UUIDs**
  - ALICE

Common rules for tape collocation are needed for T0 and T1s

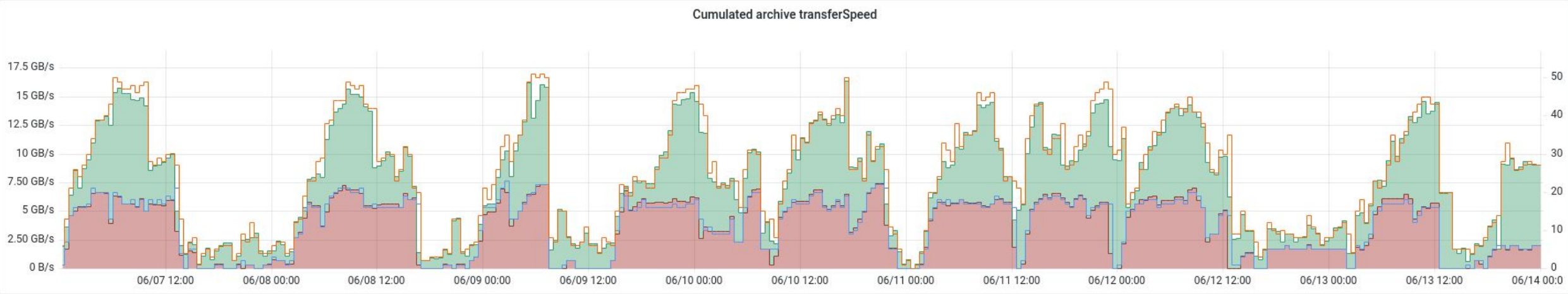
- **For example FZK file families prototype for HPSS tape cache**
- **Requires additional metadata: dataset total size, dataset file count**

We need to standardize archive metadata and work together on tape collocation at various levels

Improve tape scheduling

Additional production constraints:

- **T0 tape closely follows LHC duty cycle**



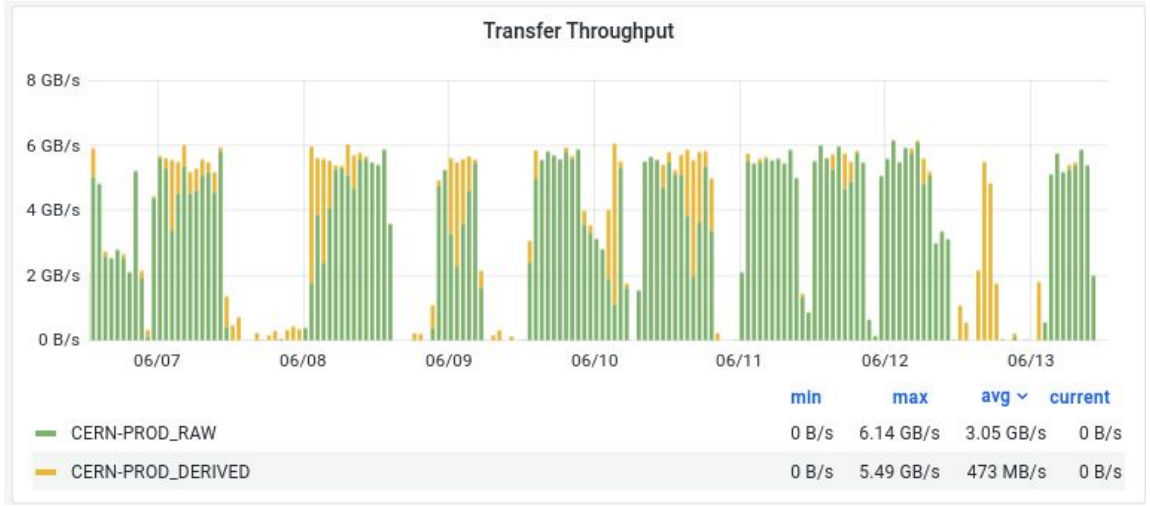
Only DAQ data must go to tape ASAP

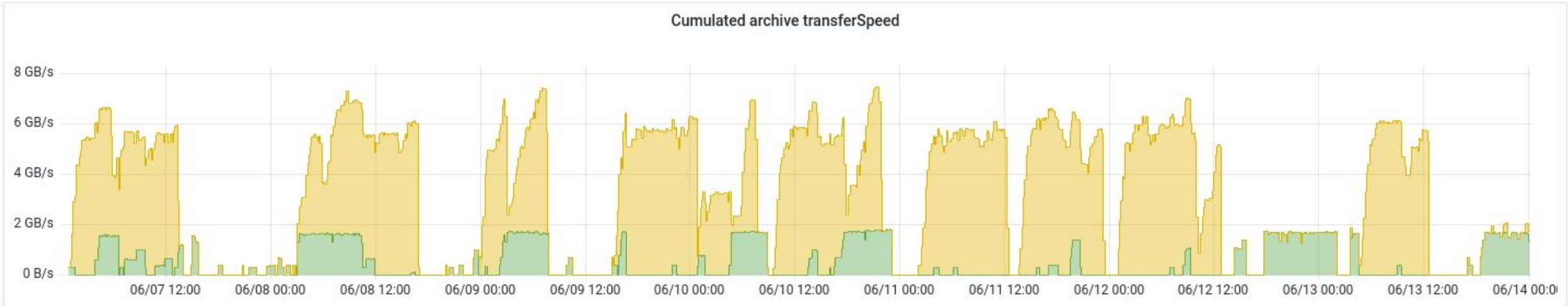
- **other traffic could wait to better spread the traffic**

Improve tape scheduling

- **ATLAS:**
  - CERN-PROD_RAW must go ASAP to tape
  - CERN-PROD_DERIVED could wait for beam dump

Would allow to move 14% of current CERN-PROD traffic outside of peak

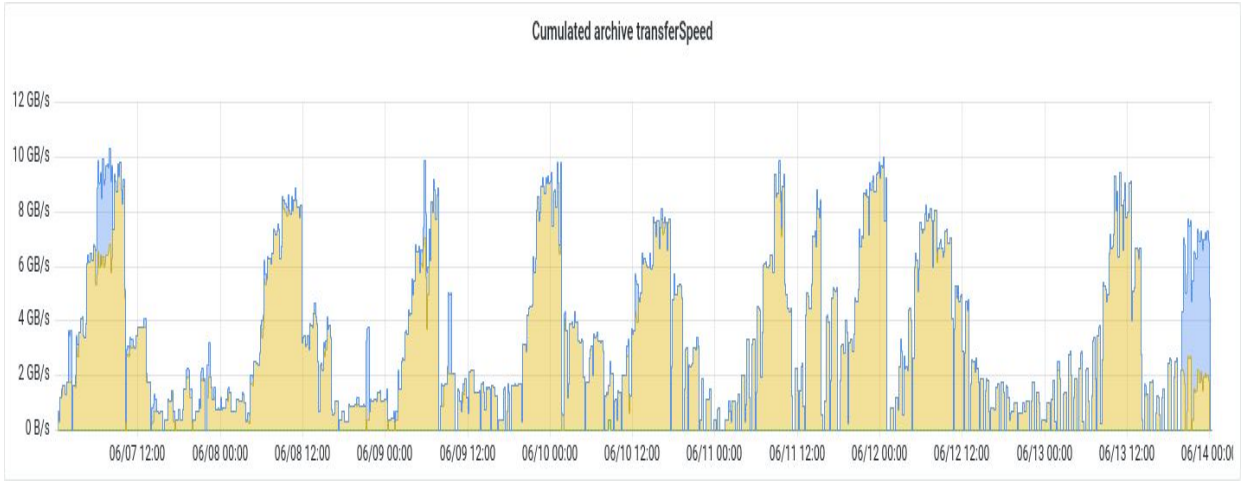


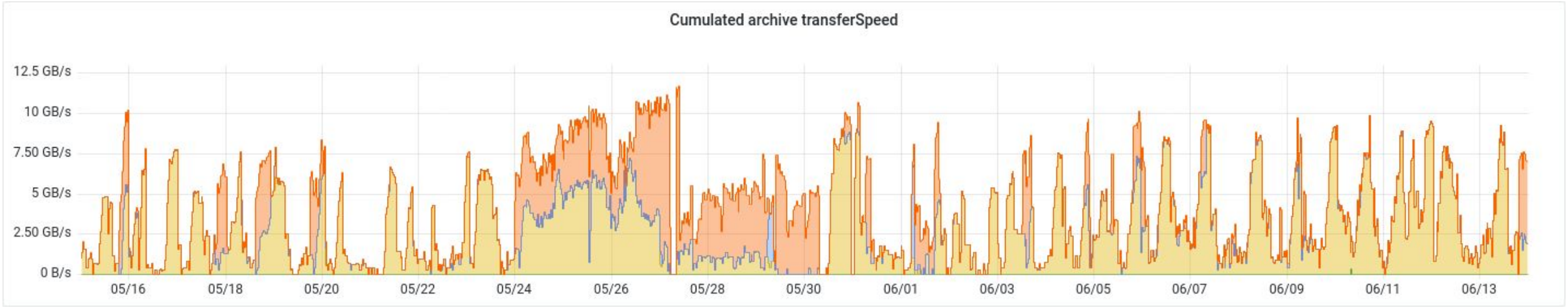


Improve tape scheduling

- **CMS:**
  - DAQ must go ASAP to tape
  - MC could wait for beam dump

Would allow to move large chunk of total T0 traffic outside of peak





Improve tape scheduling

Improving tape efficiency does not only mean increasing tape bandwidth peaks during stable beam

- **DAQ infrastructure tied with detectors**
  - Fixed max throughput for the run is defined by DAQ buffer hardware choices set at the very end of previous LS for all experiments
- **Secondary traffic is very likely to increase during the run**
  - Rebalancing less time sensitive secondary traffic could offer operational margins
  - Improve tape efficiency avoiding having most of the hardware idling when no beam

Provide some scheduling hints along with collocation metadata in archive metadata

Archive metadata proposal

- **HTTP protocol only**
  - T0 moved to HTTP only transfers for ATLAS on 230318
  - SRM going away
  - archive metadata is not part of TAPE REST API (SRM replacement)
- **CTA/dcache development agreement**
  - limit number of keys for DB:
    - **up to 4 experiment agnostic hierarchical levels for *collocation_hints***
- **Discussed with experiments DM teams**
- **FTS transparently encapsulates archive_metadata**
  - header in HTTP file transfer stream
  - **no bigger than 1kB (otherwise rejected)**

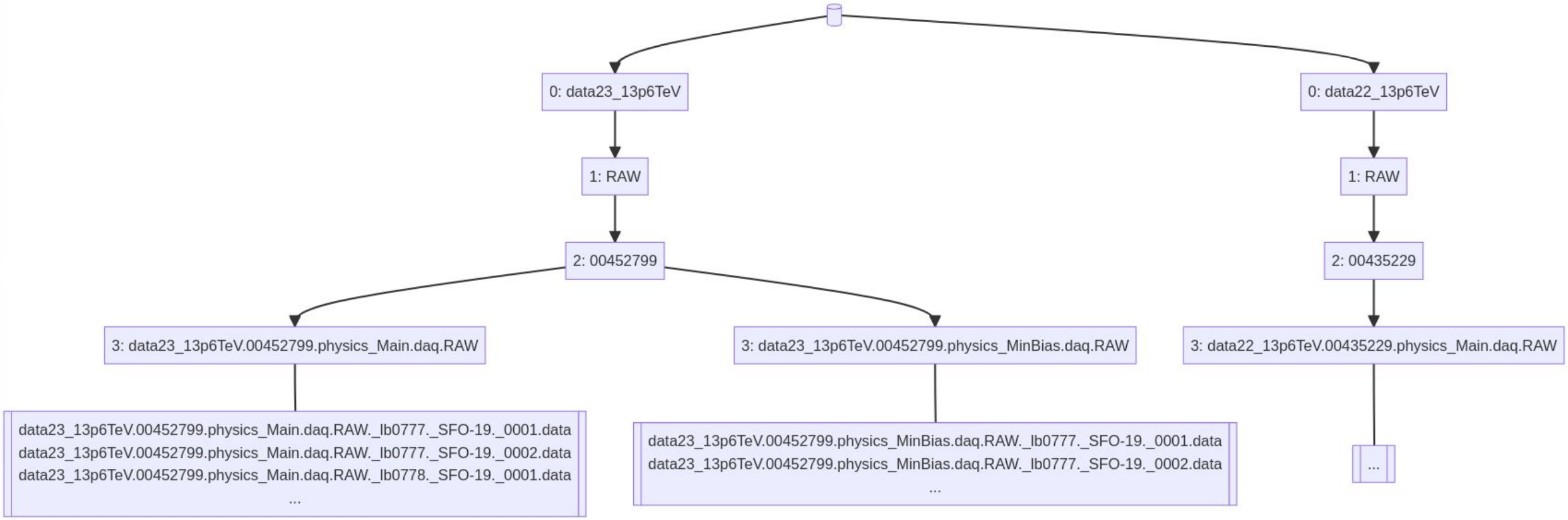
- **Archive metadata is only a hint tape sites are free to ignore**
  - common way to express archive traffic and placement constraints **from tape point of view**
    - common language for data carousel discussions
- **Everyone is working on data placement improvements**
  - CTA team starting work on new tape scheduler
  - FZK, BNL on placement improvements for HPSS

Archive metadata proposal example for DAQ file

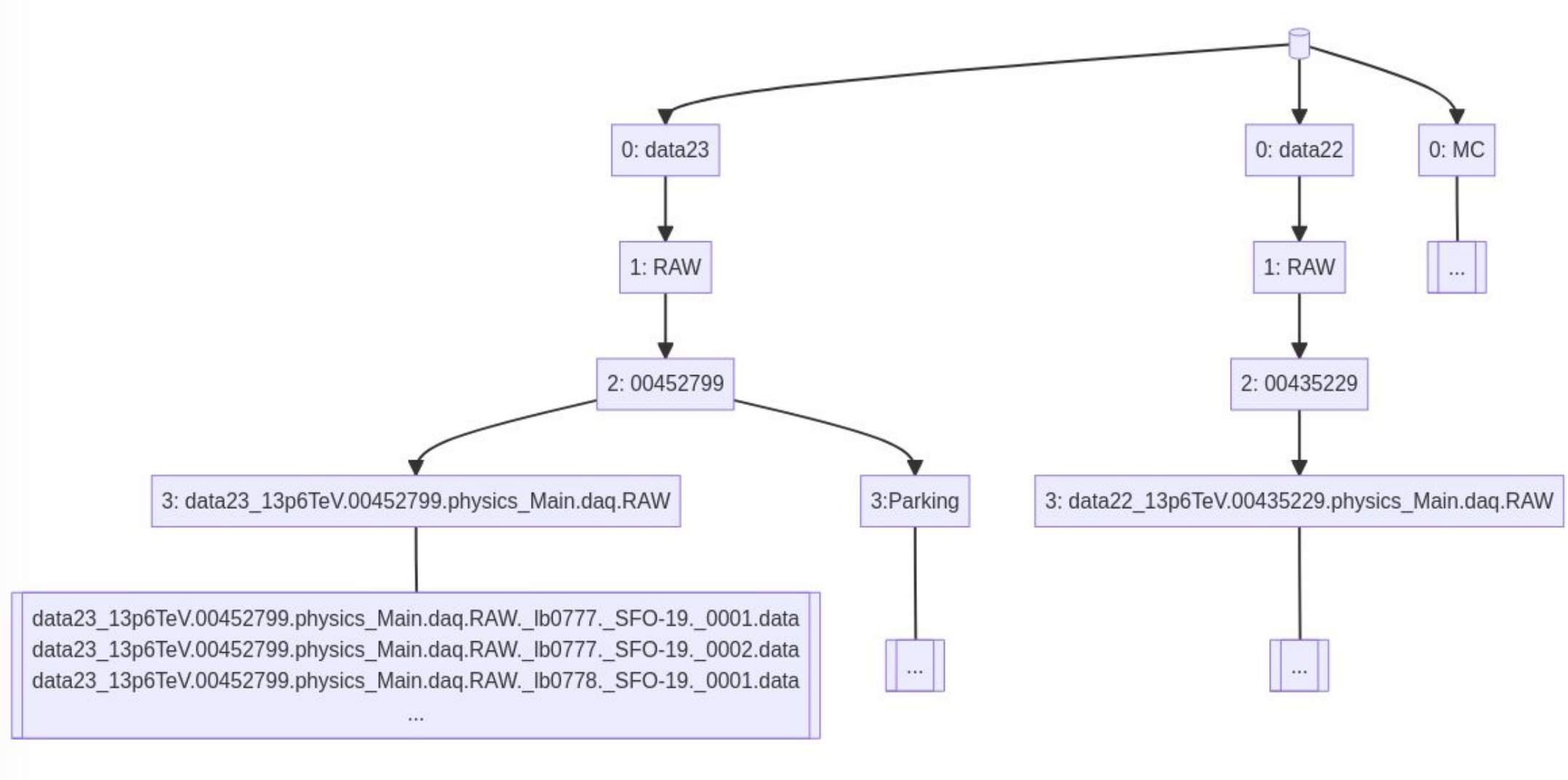
File: `data23_13p6TeV.00452799.physics_Main.daq.RAW._lb0777._SFO-19._0001.data`

```json
archive_metadata = {
    "scheduling_hints": {
        "archive_priority": "100"            # highest priority
    },
    "collocation_hints": {
        "0": "data23_13p6TeV",                                      # project
        "1": "RAW",                                                 # datatype
        "2": "00452799",                                            # runnumber
        "3": "data23_13p6TeV.00452799.physics_Main.daq.RAW",        # dataset
    },
    "optional_hints": {
        "activity": "T0 Tape",          # Tier-0/DAQ
        "3": {                          # dataset level
           "length": "19123",               # total number of files at specified level
           "bytes": "800207993184S6"        # total size of files at specified level
        }
    }
}
```

Archive metadata proposal example



Archive metadata proposal example



Archive metadata proposal

- **collocation_hints**
  - Define tree structure using tree depth as key
  - Distance between consecutive files on tape can be easily measured with a metric on the tree structure: for example using node distance in collocation tree
    - Allow to improve collocation during tape repack operations
    - Evaluate worst collocated tapes
      - for example sum of square of tree distance between 2 consecutive files

Total distance:

- **distance of 44 for 12 files on left ($2^2+2^2+2^2\ldots$)**
- **distance of 12 for 32 files on right ($0^2+2^2+0^2+0^2+\ldots$)**
- **local tape site experts can measure and later expose dataset retrieve cost**
  - per tape positioning time and transfer time, #drives, expected BW

Archive metadata proposal

- **scheduling_hints**
  - **archive_priority: "0" to "100"**
    - "0" is lowest priority, 100" is highest
    - distinct from RUCIO priority, FTS priority: *archive_* is important here
  - If bandwidth to tape is too constrained
    - exceeding allocated experiment pledge
    - sudden loss of bandwidth (tape hardware failure on site,...)
  - Allow to apply *backpressure* on archive transfers
    - protects DAQ data transfers

There is no point accepting files in tape buffer/cache if their time to tape is expected to exceed FTS *archive_timeout*

For CMS workflows it would mean potential holes on tape, consuming network throughput, filling tape buffer, tape throughput and worsening an already bad situation.

Tape site should refuse writes for lower priority transfers under bad conditions: later retried by RUCIO when situation is better: outside of DAQ peak, tape hardware issue fixed… RUCIO retry period = f(archive_priority), archive_priority could be increased for subsequent retries,...

Archive metadata proposal

- **optional_hints**
  - "*collocation_hint_key*"
    - "length": "number of files"
    - "bytes": "size in bytes"
    - For our example the metadata of the dataset is expressed
    - Allows T1s and T2s to understand how the data will fit in the tape cache, if dataset should be flushed earlier to tape, split if it is too large to be flushed together...
  - "activity"
    - for sites willing to consume it
    - suggested by Rucio but not for all experiments…

Outlook

- **CTA delivers nominal archival performance for Run3 with significant write efficiency improvements**
  - initially limited data placement features
- **Tape and protocol consolidation ongoing on the grid**
  - Opportunity to formalize and consolidate tape workflows should not be missed
- **NEXT STEP clearly oriented toward monitoring and improving data placement for tape data reads**
  - HTTP only
  - Already supported in FTS
  - **Tape sites need something simple quickly**
    - working on tape data placement improvements requires:
      - better understanding of experiment archive traffic constraints
      - gives feedback to DM teams better understanding of traffic conditions
      - work on concrete examples: archive metadata for T0 MC VS DAQ for same dataset? parking data inside DAQ?